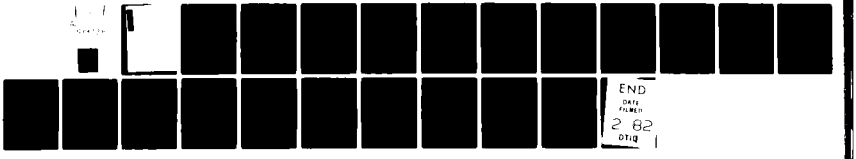
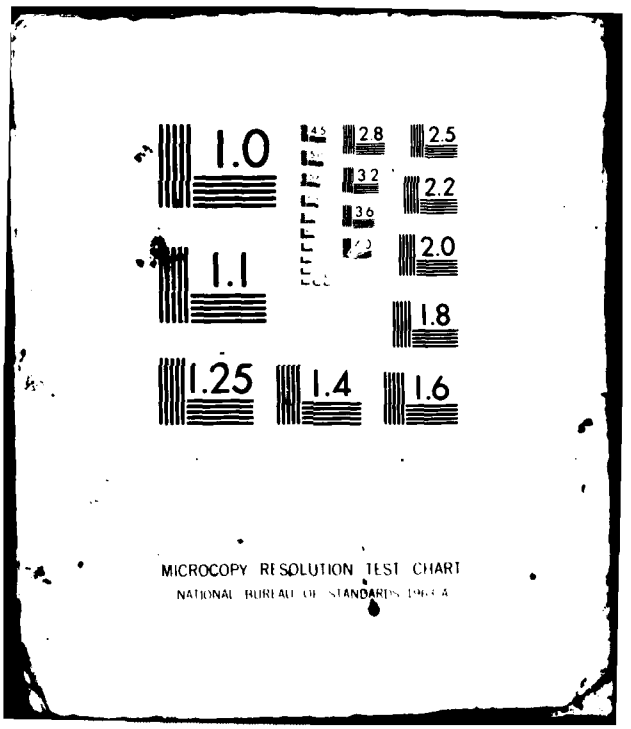
AD-A109 775 STANFORD UNIV CA DEPT OF OPERATIONS RESEARCH F/G 12/1
INDIVIDUAL VERSUS SOCIAL OPTIMIZATION IN THE ALLOCATION OF CUST--ETC(U)
JUL 81 C E BELL, S STIDHAM N00014-76-C-0418
UNCLASSIFIED TR-99 NL



END
DATE FILMED
2 82
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

INDIVIDUAL VERSUS SOCIAL OPTIMIZATION IN THE ALLOCATION OF CUSTOMERS TO ALTERNATIVE SERVERS

By

Colin E. Bell and Shaler Stidham, Jr.*

TECHNICAL REPORT NO. 99

JULY 1981

PREPARED UNDER CONTRACT

N00014-76-C-0418 (NR-047-061)

FOR THE OFFICE OF NAVAL RESEARCH

Frederick S. Hillier, Project Director

This research was supported in part by National Science Foundation Grant ECS 80-17867 Department of Operations Research, Stanford University and issued as Technical Report No. 61.

*The research of this author was partially supported by National Science Foundation Grant No. ENG 78-24420 at North Carolina State University.

DEPARTMENT OF OPERATIONS RESEARCH

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

INDIVIDUAL VERSUS SOCIAL OPTIMIZATION IN THE ALLOCATION OF CUSTOMERS TO ALTERNATIVE SERVERS

by

Colin E. Bell
Department of Management Sciences
The University of Iowa

and

Shaler Stidham, Jr.
Department of Industrial Engineering
North Carolina State University

Abstract

Customers arrive at a service area according to a Poisson process. An arriving customer must choose one of K servers without observing present congestion levels. The only available information about the k-th server is the service time distribution (with expected duration μ_k^{-1}) and the cost per unit time of waiting at the k-th server (h_k). Although service distributions may differ from server to server and need not be exponential, it is assumed that they share the same coefficient of variation. Individuals acting in self-interest induce an arrival rate pattern $(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_K)$.

In contrast, the social optimum is the arrival rate pattern $(\lambda_1^*, \lambda_2^*, \ldots, \lambda_K^*)$ which minimizes long run average cost per unit time for the entire system. The main result is that $\hat{\lambda}_k$'s and $\hat{\lambda}_k^*$'s differ systematically. Individuals overload the servers with the smallest h_k/μ_k values.

For an exponential service case with pre-emptive LIFO service an alternative charging scheme is presented which confirms that differences between individual and social optima occur precisely because individuals fail to consider the inconvenience that they cause to others.

Following the pioneering work of Naor [22], considerable attention has been devoted to comparing individual joining behavior to socially optimal joining behavior at a queuing system. The works of Naor [22], Lippman and Stidham [19], Adler and Naor [1], Knudsen [16], Knudsen and Stidham [17], Yechiali [29,30] and others all indicate that individuals acting in self-interest tend to over-congest a system relative to the social optimum. This phenomenon arises because a self-interested individual fails to consider the impact of his joining on later arrivals.

In the above models the joining decision depends on the observed current congestion level at the facility, thus inducing a state-dependent arrival rate. In contrast, our model will assume that: (1) the overall arrival rate is fixed but upon arrival customers must choose (or be assigned to) one of several alternative servers, each with its own queue. Balking, reneging, or jockeying are not permitted. (2) customers can not observe the current queue length at each server, but they are aware of the service-time distributions and waiting costs at the various servers. We will demonstrate that there are systematic differences between individual and social optimization; for a special case these differences will be shown to result from the failure of an individual to consider the impact on others of his joining a particular queue.

Since the allocations in our model are not functions of the state of the system, it is a design model, in contrast to the control models of Bell [2-6], Doshi [9], Heyman [14], Lippman and Stidham [19], Sobel [25], Winston [28] and others cited in the survey papers of Stidham and Prabhu [24] and Sobel [26]. Alternative design models for stochastic service systems have been studied by Morse [21], Hillier [15], Mangelsdorf [20], Evans [10], Kumin [18], Stidham [23] and others.

Section 1 contains preliminary results concerning the socially and individually optimal selection of the arrival rate to an isolated M/G/1 facility, with a fixed reward for each entering customer and a linear waiting cost. These results are of independent interest and also form the basis for our Lagrangian analysis of the multi-facility model. Section 2 presents the basic multi-facility model, analyzes the social-optimum and individual-optimum problems, and compares the allocations. For the main result of Section 2 (Theorem 4) we do not require exponential servers but do insist that the service-time distributions share a common coefficient of variation. A counterexample is provided for the case where the coefficients of variation are allowed to differ. Section 3 treats the special case of exponential servers in more detail. Explicit formulas are given for the socially and individually optimal allocations to each facility as functions of the overall arrival rate. We also give (in Section 4) an alternate charging scheme for inducing socially optimal behavior on the part of individuals.

1. Preliminary Results: Isolated Single-Facility Model

We assume:

(1) a Poisson arrival process with rate λ (a decision variable);

(2) a single server, who provides services of random duration S with $E(S) = \mu^{-1}$, $E(S^2) = b\mu^{-2}$; the known constant b is (necessarily) greater than or equal to 1; the coefficient of variation of S is then $(b-1)^{\frac{1}{2}}$;

(3) a fixed reward α earned for each customer who enters the system;

(4) a waiting cost h per customer per unit time in the queue and in service.

The stipulation that λ is a decision variable is satisfied, for example, in the following scenario: Arrivals to the facility are governed by a Poisson process with fixed rate Λ. An arriving customer may join or balk. The probability, p, that an arriving customer joins is a decision variable. Given a choice of p, the induced process of joining customers is Poisson with rate $\lambda = p\Lambda$.

The objective of social optimization is to choose an arrival rate λ, $0 \leq \lambda < \mu$, to minimize the long-run average net cost per unit time. (Rewards are treated as negative costs.) If arrival rate λ prevails, then the facility is an M/G/1 queue with an average of $L(\lambda)$ customers present. The Pollaczek-Khintchine formula (cf. Cox and Smith [7] or Gross and Harris [12]) yields:

$$L(\lambda) = (\lambda/\mu) + \lambda^2 b/(2\mu(\mu - \lambda)). \tag{1}$$

The social optimum λ^* is then the solution to:

$$\text{minimize } hL(\lambda) - \alpha\lambda = \lambda[(1/\mu) + \lambda b/(2\mu(\mu - \lambda)) - \alpha] \tag{2}$$

$$\text{subject to} \quad 0 \leq \lambda < \mu.$$

The objective function in (2) is differentiable and strictly convex. Denoting the derivative of $L(\cdot)$ evaluated at λ by $L'(\lambda)$, it follows immediately that an optimal solution λ^* is characterized by

$$hL'(\lambda^*) = \alpha, \quad \text{and} \quad \lambda^* > 0 \tag{3}$$

or

$$hL'(\lambda^*) = h/\mu > \alpha, \quad \text{and} \quad \lambda^* = 0. \tag{4}$$

Also from (1)

$$L'(\lambda) = \mu^{-1} + b\lambda(2\mu - \lambda)/[2\mu(\mu - \lambda)^2]. \tag{5}$$

Let $\lambda^*(\alpha)$ be the value of λ^* determined by (3) and (4) for fixed $\alpha > 0$. It follows from (3), (4), and (5) that we can represent $\lambda^*(\alpha)$ by

$$\lambda^*(\alpha) = \max\{0;\ \mu - \mu[b/(b + 2\alpha h^{-1}\mu - 2)]^{\frac{1}{2}}\}. \tag{6}$$

Note that $\lambda^*(\alpha) = 0$ for $0 < \alpha \leq h/\mu$. (There is nothing to be gained by admitting customers if the waiting cost incurred during service alone exceeds the reward.) Moreover, $\lambda^*(\alpha)$ is positive, strictly increasing, and strictly concave in $\alpha > h/\mu$, with $\lim_{\alpha \to \infty} \lambda^*(\alpha) = \mu$, as expected.

Now consider the behavior of individual customers, each seeking to maximize his expected net benefit (minimize his expected net cost). Since all customers have the same information upon arrival, it is reasonable to assume that all customers will use the same randomized strategy for deciding whether to join or balk. If λ is the arrival rate induced by this common strategy (e.g., $\lambda = p\Lambda$, where p is the probability of joining, as described above), then the facility behaves as an M/G/1 system with average wait in system, $W(\lambda)$, for a customer who enters given by

$$W(\lambda) = L(\lambda)/\lambda = \mu^{-1} + \lambda b/(2\mu(\mu - \lambda)). \tag{7}$$

The average cost of the wait in system is thus $hW(\lambda)$. Note that the rightmost expression in (7) can be used to define $W(0) = \mu^{-1}$.

In order for the common strategy of individual customers to be in equilibrium it is necessary and sufficient that the expected net benefit to a joining customer be equal to 0, if $\lambda > 0$, and less than or equal to 0, if $\lambda = 0$. Otherwise, an individual customer would have an incentive to deviate from the common strategy either by joining with probability one if the expected net benefit is positive, or by balking with probability one if the expected net benefit is negative. Denoting the equilibrium

(individually optimal) value of λ by $\hat{\lambda}$, the necessary and sufficient conditions for equilibrium are therefore

$$hW(\hat{\lambda}) = \alpha, \quad \text{and} \quad \hat{\lambda} > 0 \tag{8}$$

or

$$hW(\hat{\lambda}) = h/\mu \geq \alpha, \quad \text{and} \quad \hat{\lambda} = 0. \tag{9}$$

Thus, although there is no overall objective to be minimized, the requirements (8), (9) for a solution based on individual optimization are analogous to the requirements (3), (4) for a social optimum. Denoting by $\hat{\lambda}(\alpha)$ the value of $\hat{\lambda}$ determined by (8) and (9) for fixed α, it follows from (7), (8), and (9) that

$$\hat{\lambda}(\alpha) = \max\{0;\ \mu - \mu[b/(b + 2\alpha h^{-1}\mu - 2)]\}. \tag{10}$$

Like $\lambda^*(\alpha)$, $\hat{\lambda}(\alpha) = 0$ for $0 < \alpha \leq h/\mu$, and $\hat{\lambda}(\alpha)$ is positive, strictly increasing, and strictly concave in $\alpha > h/\mu$, with $\lim_{\alpha \to \infty} \hat{\lambda}(\alpha) = \mu$.

Having determined conditions satisfied by the social optimum λ^* and the individual optimum $\hat{\lambda}$, we are in a position to compare the two solutions.

Lemma 1. For $0 < \alpha \leq h/\mu$, $\lambda^*(\alpha) = \hat{\lambda}(\alpha) = 0$. For $\alpha > h/\mu$, $\lambda^*(\alpha) < \hat{\lambda}(\alpha)$, that is, the individual optimum solution allows more customers to enter than is socially optimal.

Proof. The first assertion is obvious. The second assertion can be proved directly by comparing (6) and (9). The following alternate proof, however, is perhaps more instructive and has wider applicability.

The objective function in the social-optimum problem takes the form:

$$C(\lambda) = \lambda(hW(\lambda) - \alpha).$$

Since $W(\lambda)$ is strictly increasing in λ, we see that $C(\lambda) < 0$ for $0 \leq \lambda < \hat{\lambda}(\alpha)$, $C(\hat{\lambda}(\alpha)) = 0$, and $C(\lambda) > 0$ for $\lambda > \hat{\lambda}(\alpha)$. Since $\lambda^*(\alpha)$ minimizes $C(\lambda)$, it follows that $\lambda^*(\alpha) < \hat{\lambda}(\alpha)$.

Note that the proof of Lemma 1 required only that $W(\lambda)$ be strictly increasing in $\lambda > 0$, with $W(0) = \mu^{-1}$. Hence Lemma 1 is valid for any queuing system for which these properties hold. In fact one expects that these properties would hold in general in queuing systems in which $W(\lambda)$ is well defined as a long-run average within a suitable class of arrival processes parameterized by the mean arrival rate λ, as is the case with Poisson processes.

2. Basic Results for Multi-Facility Model

Now we assume:

(1) a Poisson arrival process with fixed rate $\Lambda > 0$;

(2) K servers; server k provides services of random duration S_k with $E(S_k) = \mu_k^{-1}$, $E(S_k^2) = b\mu_k^{-2}$; $\sum_{k=1}^{K} \mu_k > \Lambda$;

(3) arrival rate λ_k at server k (a decision variable); an arrival to the system is assigned to server k with probability λ_k/Λ;

(4) a waiting cost h_k per customer per unit time in the queue and in service at server k;

(5) servers numbered so that $h_1\mu_1^{-1} \leq h_2\mu_2^{-1} \leq \cdots \leq h_K\mu_K^{-1}$.

Two possible scenarios which might make the assumption of a common coefficient of variation for all service distributions plausible are: (1) the nature of the service function is such that it is reasonable to assume that all service distributions come from a parametric family with this assumption, or (2) the servers represent processors which process "work" in a deterministic fashion at a rate μ_k work units per unit time;

customers require a random duration of a server's attention because they present the server with a random number of work units; customer work unit requirements have expected value 1 and variance (b-1), coefficient of variation $(b-1)^{\frac{1}{2}}$.

The social optimization problem is one of finding an allocation, $\lambda_1^*, \lambda_2^*, \ldots, \lambda_K^*$, of customers to servers so as to minimize the long-run average cost per unit time for the entire system. If arrival rate λ_k prevails at server k, then server k is an M/G/1 queuing system with an average of $L_k(\lambda_k)$ customers present, where $L_k(\lambda_k)$ is given by formula (1) with λ and μ replaced by λ_k and μ_k respectively. The social optimum $(\lambda_1^*, \lambda_2^*, \ldots, \lambda_K^*)$ is then the solution to:

$$\begin{aligned} &\text{minimize} \sum_{k=1}^{K} h_k L_k(\lambda_k) \\ &\text{subject to} \sum_{k=1}^{K} \lambda_k = \Lambda \\ &\qquad 0 \le \lambda_k < \mu_k, \; k = 1,2,\ldots,K. \end{aligned} \tag{11}$$

It can readily be demonstrated (e.g., using generalized Lagrange multipliers) that an optimal solution to (11) satisfies equations of the form (3) and (4) for each k and some value of α such that $\sum_{k=1}^{K} \lambda_k^* = \Lambda$. Thus (11) can be solved by a one-dimensional search over $\alpha > 0$, with $\lambda_k^*(\alpha)$ defined (as in the single-facility model) by

$$\lambda_k^*(\alpha) = \max\{0, \; \mu_k - \mu_k[b/(b + 2\alpha h_k^{-1}\mu_k - 2)]^{\frac{1}{2}}\}. \tag{12}$$

Since $\sum_{k=1}^{K} \lambda_k^*(\alpha)$ is strictly increasing in $\alpha > h_1\mu_1^{-1}$, we need only find the unique α^* with $\sum_{k=1}^{K} \lambda_k^*(\alpha^*) = \Lambda$. Since $\lambda_k^*(\alpha) > 0$ iff $\alpha > h_k\mu_k^{-1}$, it is clear that $\lambda_k^* > 0$ implies $\lambda_j^* > 0$ for $j < k$. Only the highest numbered servers (if any) are not used at all.

The individual optimization problem is one of finding an equilibrium common-strategy arrival-rate vector $(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_K)$. The average wait in system k for a customer who enters, denoted $W_k(\lambda_k)$, is given by formula (7), with λ and μ replaced by λ_k and μ_k, respectively. By analogy with the single-facility problem, in order for the arrival-rate vector to be in equilibrium it is necessary and sufficient that (i) the average cost of waiting, $h_k W_k(\hat{\lambda}_k)$, be equal at all servers k that are used (i.e., with $\hat{\lambda}_k > 0$) and (ii) any server k with $\hat{\lambda}_k = 0$ cannot have $h_k \mu_k^{-1}$ less than the common average waiting cost at used servers. Similar necessary and sufficient conditions for equilibrium were established by Dafermos [8], Hall and Peterson [13], Wardrop [27], and other authors cited in Florian [11] in the context of traffic flows in a network. It follows that an equilibrium solution satisfies equations of the form (8) and (9) for each k and some value of α such that $\sum_{k=1}^{K} \hat{\lambda}_k(\alpha) = \Lambda$. Again, an equilibrium solution can be found by a one-dimensional search over $\alpha > 0$, with $\hat{\lambda}_k(\alpha)$ defined by:

$$\hat{\lambda}_k(\alpha) = \max\{0,\ \mu_k - \mu_k[b/(b + 2\alpha h_k^{-1}\mu_k - 2)]\}. \tag{13}$$

Since $\sum_{k=1}^{K} \hat{\lambda}_k(\alpha)$ is strictly increasing in $\alpha > h_1 \mu_1^{-1}$, we need only find the unique $\hat{\alpha}$ with $\sum_{k=1}^{K} \hat{\lambda}_k(\hat{\alpha}) = \Lambda$. Since $\hat{\lambda}_k(\alpha) > 0$ iff $\alpha > h_k \mu_k^{-1}$, it is clear that $\hat{\lambda}_k > 0$ implies $\hat{\lambda}_j > 0$ for $j < k$.

We now compare the social optimum and individual optimum solutions for the multi-facility problem. We first prove Lemma 2, which establishes that $\lambda_k^* = 0$ implies $\hat{\lambda}_k = 0$: the social optimum makes use of at least as many servers as the individual solution.

Lemma 2. $\alpha^* > \hat{\alpha}$.

Proof. Suppose $\alpha^* \leq \hat{\alpha}$. Then from (12) and (13), $\hat{\lambda}_k = 0$ implies $\lambda_k^* = 0$. On the other hand, if $\hat{\lambda}_k > 0$, then by Lemma 1 $\hat{\lambda}_k = \hat{\lambda}_k(\hat{\alpha}) \geq \hat{\lambda}_k(\alpha^*) > \lambda_k^*(\alpha^*) = \lambda_k^*$. Hence $\sum_{k=1}^{K} (\hat{\lambda}_k - \lambda_k^*)$ is positive, since it contains at least one positive term and no negative terms. This contradicts the requirement that $\sum_{k=1}^{K} \hat{\lambda}_k = \sum_{k=1}^{K} \lambda_k^* = \Lambda$. Thus $\alpha^* > \hat{\alpha}$.

For each $k = 1,2,\ldots,K$, define $\rho_k^* := \lambda_k^*/\mu_k$ and $\hat{\rho}_k := \hat{\lambda}_k/\mu_k$. Dividing both sides of (12) and (13) by μ_k yields the following result, which is used in the proof of the main theorem.

Lemma 3. For $i < j$,

(i) $\lambda_i^* > 0 \Rightarrow \rho_i^* \geq \rho_j^*$ with equality only if $h_i\mu_i^{-1} = h_j\mu_j^{-1}$;

(ii) $\hat{\lambda}_i > 0 \Rightarrow \hat{\rho}_i \geq \hat{\rho}_j$ with equality only if $h_i\mu_i^{-1} = h_j\mu_j^{-1}$.

We are now ready for the main theorem, which shows that individual optimization systematically overloads the lower numbered servers (those with smallest h_k/μ_k).

Theorem 4. If $\lambda_i^* \geq \hat{\lambda}_i$, then $\lambda_j^* \geq \hat{\lambda}_j$ for all $j > i$.

Proof. Assume $i < j$. First $\lambda_i^* = 0$ and $\lambda_i^* \geq \hat{\lambda}_i$ imply $\hat{\lambda}_i = 0$ and hence $\lambda_j^* = \hat{\lambda}_j = 0$ by Lemma 3 and thus the theorem holds. Now suppose $\lambda_i^* > 0$ and $\lambda_i^* \geq \hat{\lambda}_i$, but $\lambda_j^* < \hat{\lambda}_j$. Then Lemmas 2 and 3 imply all of λ_i^*, $\hat{\lambda}_i$, λ_j^*, $\hat{\lambda}_j$ are positive. From (3) and (8)

$$h_i L_i'(\lambda_i^*) = \alpha^* = h_j L_j'(\lambda_j^*) \tag{14}$$

$$h_i W_i(\hat{\lambda}_i) = \hat{\alpha} = h_j W_j(\hat{\lambda}_j). \tag{15}$$

Define $\rho := \lambda/\mu$ and

$$f(\rho) := \rho b/(2(1-\rho)). \tag{16}$$

Note that $f(\rho)$ is differentiable, non-negative, strictly increasing, and strictly convex in $0 \leq \rho < 1$. We shall also need the following property of $f(\cdot)$, which is easily verified from (16) by differentiation:

$$\rho f'(\rho)/f(\rho) \text{ is strictly increasing in } 0 \leq \rho < 1. \tag{17}$$

It follows from (1) and (7) that $L(\lambda) = \rho(1+f(\rho))$, $L'(\lambda) = \mu^{-1}(1+f(\rho)+\rho f'(\rho))$, and $W(\lambda) = \mu^{-1}(1+f(\rho))$. Hence (14) and (15) can be rewritten in equivalent form:

$$(h_i\mu_i^{-1})(1+f(\rho_i^*)+\rho_i^* f(\rho_i^*)) = \alpha^* = (h_j\mu_j^{-1})(1+f(\rho_j^*)+\rho_j^* f(\rho_j^*)) \tag{18}$$

$$(h_i\mu_i^{-1})(1+f(\hat{\rho}_i)) = \hat{\alpha} = (h_j\mu_j^{-1})(1+f(\hat{\rho}_j)). \tag{19}$$

By the hypotheses, $\rho_i^* \geq \hat{\rho}_i$ and $\rho_j^* < \hat{\rho}_j$. It follows by subtracting (19) from (18) and using the fact that $f(\cdot)$ is strictly increasing and strictly convex that

$$(h_i\mu_i^{-1})\hat{\rho}_i f'(\hat{\rho}_i) \leq \alpha^* - \hat{\alpha} < (h_j\mu_j^{-1})\hat{\rho}_j f'(\hat{\rho}_j). \tag{20}$$

But (19) and the fact that $h_i\mu_i^{-1} \leq h_j\mu_j^{-1}$ imply that

$$(h_i\mu_i^{-1})f(\hat{\rho}_i) \geq (h_j\mu_j^{-1})f(\hat{\rho}_j). \tag{21}$$

Dividing (20) by (21) yields

$$\hat{\rho}_i f'(\hat{\rho}_i)/f(\hat{\rho}_i) < \hat{\rho}_j f'(\hat{\rho}_j)/f(\hat{\rho}_j),$$

from which it follows, using property (17), that $\hat{\rho}_i < \hat{\rho}_j$, which is a contradiction of Lemma 3. Therefore, it cannot be true that both

$\rho_i^* \geq \hat{\rho}_i$ and $\rho_j^* < \hat{\rho}_j$. We conclude that $\rho_i^* \geq \hat{\rho}_i$ implies $\rho_j^* \geq \hat{\rho}_j$, the desired result.

Note that $f(\rho)$ as defined by (16) equals the average waiting time in the queue, expressed in units of mean service time. Thus Theorem 4 is valid for any multi-facility queuing system in which the average waiting time at each facility k, normalized in this way, is a function only of the traffic intensity, λ_k/μ_k, at that facility, the function is the same for all facilities, and it has the properties referred to in the proof of the theorem: strictly increasing, strictly convex, and property (17). Note also that in order for these properties to hold it suffices for f to be strictly increasing and log convex.

Theorem 4 assumes that all service time distributions share a common coefficient of variation and demonstrates that individuals overload servers k with low h_k/μ_k. If these coefficients of variation are allowed to differ from server to server one might suspect a tendency to shy away from servers with a particularly high coefficient of variation. This tendency provides our counter-example.

Let server 1 have $\mu_1 = 1 + 10^{-8}$, $h_1 = 1$, $b_1 = 2$ (e.g., exponential service) and let server 2 have $\mu_2 = 1$, $h_2 = 1$, $b_2 = 1$ (constant service) so that $h_1/\mu_1 < h_2/\mu_2$. Using (12) and (13), λ_1^* and $\hat{\lambda}_1$ can be found to be .435 and .414 respectively. Thus $(\lambda_1^* - \lambda_1) = .021$ while $(\lambda_2^* - \hat{\lambda}_2) = -.021$, contradicting Theorem 4.

3. Further Results for Exponential Servers

Throughout this section we assume an exponential service-time distribution. In this case $b = 2$ and the expressions for $\lambda_i^*(\alpha)$ and $\hat{\lambda}_i(\alpha)$ become, respectively,

$$\lambda_i^*(\alpha) = \max\{0,\ \mu_i - (h_i\mu_i/\alpha)^{1/2}\} \tag{22}$$

$$\hat{\lambda}_i(\alpha) = \max\{0,\ \mu_i - h_i/\alpha\}. \tag{23}$$

To find the social optimum allocation, $\lambda_1^*, \lambda_2^*, \ldots, \lambda_K^*$, for a given system arrival rate Λ, we know from the analysis in section 2 that it suffices to find the unique $\alpha = \alpha^*$ such that $\sum_{i=1}^{K} \lambda_i^*(\alpha^*) = \Lambda$. Similarly, to find the equilibrium allocation, $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_K$, it suffices to find the unique $\alpha = \hat{\alpha}$ such that $\sum_{i=1}^{K} \hat{\lambda}_i(\hat{\alpha}) = \Lambda$.

Define $k^* := \max\{k | \alpha^* > h_k\mu_k^{-1}\}$, $\hat{k} := \max\{k | \hat{\alpha} > h_k\mu_k^{-1}\}$. Then the open facilities in the social (individual) optimum allocation are facilities 1 through $k^*(\hat{k})$, where $k^* \geq \hat{k}$ by Lemma 2. It follows from (22) and (23) that

$$\sum_{i=1}^{k^*} (\mu_i - (h_i\mu_i/\alpha^*)^{1/2}) = \Lambda,$$

$$\sum_{i=1}^{\hat{k}} (\mu_i - h_i/\hat{\alpha}) = \Lambda,$$

from which we see that

$$(1/\alpha^*)^{1/2} = \left[\sum_{i=1}^{k^*} \mu_i - \Lambda\right] \left(\sum_{i=1}^{k^*} (h_i\mu_i)^{1/2}\right)^{-1} \tag{24}$$

$$1/\hat{\alpha} = \left[\sum_{i=1}^{\hat{k}} \mu_i - \Lambda\right] \left(\sum_{i=1}^{\hat{k}} h_i\right)^{-1} \tag{25}$$

Now k^* is the unique k such that $h_k\mu_k^{-1} < \alpha^* \leq h_{k+1}\mu_{k+1}^{-1}$ (with $h_{k+1}\mu_{k+1}^{-1}$ replaced by ∞ when $k = K$). Similarly, $\hat{k}$ is the unique k such that $h_k\mu_k^{-1} < \hat{\alpha} \leq h_{k+1}\mu_{k+1}^{-1}$. Combining these inequalities with (24) and (25) yields the following characterizations:

$$k^* = k \quad \text{iff} \quad \sum_{i=1}^{k} \mu_i - (h_k^{-1}\mu_k)^{\frac{1}{2}} \sum_{i=1}^{k} (h_i\mu_i)^{\frac{1}{2}} < \Lambda \leq \sum_{i=1}^{k} \mu_i - (h_{k+1}^{-1}\mu_{k+1})^{\frac{1}{2}} \sum_{i=1}^{k} (h_i\mu_i)^{\frac{1}{2}} \tag{26}$$

and

$$\hat{k} = k \quad \text{iff} \quad \sum_{i=1}^{k} \mu_i - (h_k^{-1}\mu_k) \sum_{i=1}^{k} h_i < \Lambda \leq \sum_{i=1}^{k} \mu_i - (h_{k+1}^{-1}\mu_{k+1}) \sum_{i=1}^{k} h_i. \tag{27}$$

Now define $r_k := \sum_{i=1}^{k} [\mu_i - (h_k^{-1}\mu_k h_i \mu_i)^{\frac{1}{2}}]$, $s_k := \sum_{i=1}^{k} [\mu_i - (h_k^{-1}\mu_k h_i\mu_i)]$, $k = 1,2,\ldots,K$, and $r_{K+1} := s_{K+1} := \sum_{i=1}^{K} \mu_i$. (Note that $r_1 = s_1 = 0$.) It follows from (22)-(27) that the socially optimal and individually optimal allocations are given by the following explicit formulae:

$$\lambda_j^* = \begin{cases} 0, \text{ if } 0 < \Lambda \leq r_j \\ \mu_j - \left[(h_j\mu_j)^{\frac{1}{2}} \Big/ \sum_{i=1}^{k} (h_i\mu_i)^{\frac{1}{2}}\right] \cdot \left(\sum_{i=1}^{k} \mu_i - \Lambda\right), \text{ if } r_k < \Lambda \leq r_{k+1}, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad j \leq k \leq K \end{cases} \tag{28}$$

$$\hat{\lambda}_j = \begin{cases} 0, \text{ if } 0 < \Lambda \leq s_j \\ \mu_j - \left[h_j \Big/ \sum_{i=1}^{k} h_i\right] \cdot \left(\sum_{i=1}^{k} \mu_i - \Lambda\right), \text{ if } s_k < \Lambda \leq s_{k+1}, \; j \leq k \leq K \end{cases} \tag{29}$$

These formulae reveal that in both the socially optimal and the individually optimal allocations, each facility gets a share, $\mu_j - \lambda_j$, of the total excess service capacity, $\sum_{i=1}^{K} \mu_i - \Lambda$, of all open facilities. In the socially optimal allocation, the share given to facility j is proportional

to $(h_j\mu_j)^{\frac{1}{2}}$, whereas in the individually optimal allocation it is proportional to h_j. Note also that λ_j^* and $\hat{\lambda}_j$ are both piecewise-linear in Λ, $0 < \Lambda < \sum_{i=1}^{K} \mu_i$, and concave over the range where they are positive.

4. An Alternative Charging Scheme for the Exponential Service Case

Throughout this section we again assume exponential service. Although we have been unable to derive analogous results for other cases, results in this section demonstrate explicitly that not being required to consider others causes individual behavior to fail to attain a social optimum.

Since $b = 2$ for exponential service, (1) can be simplified and the relationship $L = \lambda W$ used to write

$$W_k(\lambda_k) = 1/(\mu_k - \lambda_k) \tag{30}$$

and (5) can be rewritten as

$$L_k'(\lambda_k) = \mu_k/(\mu_k - \lambda_k)^2. \tag{31}$$

For individual optimization all $h_k W_k(\hat{\lambda}_k)$ for $\hat{\lambda}_k > 0$ are equated and for social optimization all $h_k L_k'(\lambda_k^*)$ for $\lambda_k^* > 0$ are equated.

None of our results to this point have required that the service discipline be specified. For clarity we assume that each server provides pre-emptive LIFO service. Under this assumption an arriving customer at server k remains in the system for one M/M/1 busy period and pays h_k per unit time while there. With LIFO service he inconveniences those customers who are already in the system and forces them to wait through one additional busy period; however, he is not charged in any way for inconveniencing others.

As an alternative, assume that an arriving customer must pay h_k per unit time for himself and for each of the other customers whom he inconveniences. Then his average cost is

$$h_k(\mu_k - \lambda_k)^{-1}[1 + \lambda_k(\mu_k - \lambda_k)^{-1}] \tag{32}$$

where $(\mu_k - \lambda_k)^{-1}$ is the average busy period length, $\lambda_k(\mu_k - \lambda_k)^{-1}$ is the average number of other customers present when he arrives, and the term in brackets in (32) is the average number of customers inconvenienced including himself. But (32) simplifies to $h_k\mu_k(\mu_k - \lambda_k)^{-2}$. With this revised charging scheme an equilibrium arrival rate vector $(\bar{\lambda}_1, \bar{\lambda}_2, \ldots, \bar{\lambda}_K)$ would be established equating $h_k\mu_k(\mu_k - \bar{\lambda}_k)^{-2}$ for all k with $\bar{\lambda}_k > 0$. But these are exactly the conditions for a social optimum. Thus (in this context) individuals will act in a socially optimal way when they are forced to explicitly pay for the inconvenience caused to others.

References

1. Adler, I. and Naor, P., "Social Optimization versus Self-Optimization in Waiting Lines," Technical Report No. 4, Department of Operations Research, Stanford University, Stanford, Calif. (1969).

2. Bell, C. E., "Characterization and Computation of Optimal Policies for Operating an M/G/1 Queuing System with Removable Server," Operations Research 19, 208-218 (1971).

3. Bell, C. E., "Optimal Operation of an M/G/1 Priority Queue with Removable Server," Operations Research 21, 1281-1290 (1973).

4. Bell, C. E., "Optimal Operation of an M/M/2 Queue with Removable Servers," Operations Research 28, 1189-1204 (1980).

5. Bell, C. E., "Rational Policies in an M/M/c Queue with Removable Servers," Working Paper, College of Business Administration, University of Tennessee (1979).

6. Bell, C. E., "Turning Off a Server with Customers Present: Is This Any Way to Run an M/M/c Queue with Removable Servers?" Operations Research 23, 571-574 (1975).

7. Cox, D. R., and W. L. Smith, Queues, Methuen Monograph (1961).

8. Dafermos, S. C., "The Traffic Assignment Problem for Multiclass-User Transportation Networks, Transportation Science 6, 73-87 (1972).

9. Doshi, B. T. (1977) "Continuous-Time Control of the Arrival Process in an M/G/1 Queue," Stochastic Processes and Their Applications 5 (to appear).

10. Evans, R. V., "Design of Discrete Parameter Markov Systems," Tech. Memo. No. 127, Operations Research Department, Case Western Reserve University (1968).

11. Florian, M. A. (ed.), Traffic Equilibrium Methods, Springer-Verlag, New York (1976).

12. Gross, D. and C. M. Harris, Fundamentals of Queuing Theory, John Wiley & Sons, New York (1974).

13. Hall, M. A. and E. L. Peterson, "Highway Traffic Equilibria Analysed via Geometric Programming," in Traffic Equilibrium Methods, pp. 119-131, M. A. Florian (ed.), Springer-Verlag, New York (1976).

14. Heyman, D. P., "Optimal Operating Policies for M/G/1 Queuing Systems," Operations Research 16, 362-382 (1968).

15. Hillier, F., "Economic Models for Industrial Waiting Line Problems," Management Science 10, 119-130 (1963).

16. Knudsen, N. C., "Individual and Social Optimization in a Multiserver Queue with a General Cost-Benefit Structure," Econometrica 40, 515-528 (1972).

17. Knudsen, N. C. and S. Stidham, "Individual and Social Optimization in Birth-Death Congestion Systems with a General Cost-Benefit Structure," Technical Report No. 43, Operations Research Department, Stanford University (1976).

18. Kumin, H., "The Design of Markovian Congestion Systems," Tech. Memo. No. 115, Operations Research Department, Case Western Reserve University (1968).

19. Lippman, S. A. and S. Stidham, "Individual versus Social Optimization in Exponential Congestion Systems," Operations Research 25, 233-247 (1977).

20. Mangelsdorf, T. M., "Waiting Line Theory Applied to Manufacturing Problems," S. M. Thesis, M.I.T. (1965). Reprinted in Analysis of Industrial Operations, Bowman, E. H. and R. B. Fetter (eds.) Richard D. Irwin, Homewood, Ill. (1959).

21. Morse, R., Queues, Inventories and Maintenance, Wiley, New York (1958).

22. Naor, P., "On the Regulation of Queue Size by Levying Tolls," Econometrica 37, 15-24 (1969).

23. Stidham, S., "On the Optimality of Single-Server Queuing Systems," Operations Research 18, pp. 708-732 (1970).

24. Stidham, S. and N. U. Prabhu (1974) "Optimal Control of Queuing Systems," Mathematical Methods in Queuing Theory, Lecture Notes in Economics and Mathematical Systems 98, 263-294, Berlin-Heidelberg-New York, Springer.

25. Sobel, M. J., "Optimal Average Cost Policy for a Queue with Start-Up and Shut-Down Costs," Operations Research 17, 145-162, 1969.

26. Sobel, M. J., "Optimal Operation of Queues," Proceedings of a Conference on Mathematical Methods in Queuing Theory, New York, Springer-Verlag (1974).

27. Wardrop, J. G., "Some Theoretical Aspects of Road Traffic Research," Proc. Int. Civil Engineers, Part II(1), pp. 325-378 (1952).

28. Winston, W. (1977) "Optimality of Monotonic Policies for Multiple-Server Exponential Queuing Systems with State-Dependent Arrival Rates," Graduate School of Business, Indiana University.

29. Yechiali, U., "Customers' Optimal Joining Rules for the GI/M/s Queue," Management Science 18, 434-443 (1972).

30. Yechiali, U., "On Optimal Balking Rules and Toll Charges in a GI/M/1 Queuing Process," Operations Research 19, 349-370 (1971).

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: 99	2. GOVT ACCESSION NO.: AD-A109 775
	3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): INDIVIDUAL VERSUS SOCIAL OPTIMIZATION IN THE ALLOCATION OF CUSTOMERS TO ALTERNATIVE SERVERS	5. TYPE OF REPORT & PERIOD COVERED: TECHNICAL REPORT
	6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): COLIN E. BELL and SHALER STIDHAM, JR.	8. CONTRACT OR GRANT NUMBER(s): N00014-76-C-0418
9. PERFORMING ORGANIZATION NAME AND ADDRESS: OPERATIONS RESEARCH PROGRAM -ONR DEPARTMENT OF OPERATIONS RESEARCH STANFORD UNIVERSITY, STANFORD, CALIF.	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: (NR-047-061)
11. CONTROLLING OFFICE NAME AND ADDRESS: OFFICE OF NAVAL RESEARCH OPERATIONS RESEARCH PROGRAM CODE 434 ARLINGTON, VA. 22217	12. REPORT DATE: JULY 1981
	13. NUMBER OF PAGES: 17
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office):	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

This document has been approved for public release and sale; its distribution is unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Also issued as Technical Report No. 61 - Dept. of Operations Research Stanford University under NSF GRANT ECS80-17867.

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Queueing Theory, Control of queues, Individual optimization

Social optimization, Heterogeneous servers

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

PLEASE SEE OTHER SIDE

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

TECHNICAL REPORT NO. 99 - Authors: Colin E. Bell and Shaler Stidham, Jr.

Customers arrive at a service area according to a Poisson process. An arriving customer must choose one of K servers without observing present congestion levels. The only available information about the k-th server is the service time distribution (with expected duration μ_k^{-1}) and the cost per unit time of waiting at the k-th server (h_k). Although service distributions may differ from server to server and need not be exponential, it is assumed that they share the same coefficient of variation. Individuals acting in self-interest induce an arrival rate pattern $(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_K)$.

In contrast, the social optimum is the arrival rate pattern $(\lambda_1^*, \lambda_2^*, \ldots, \lambda_K^*)$ which minimizes long run average cost per unit time for the entire system. The main result is that $\hat{\lambda}_k$'s and λ_k^*'s differ systematically. Individuals overload the servers with the smallest h_k/μ_k values.

For an exponential service case with pre-emptive LIFO service an alternative charging scheme is presented which confirms that differences between individual and social optima occur precisely because individuals fail to consider the inconvenience that they cause to others.